ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

Privacy Preserving Private Frequent Itemset Mining via Smart Splitting

A.Monisha [1], B.S.Sangeetha[2]

M.Phil Student, Department of Computer Science, Shri Sakthikailaash Women's College, Salem, Tamilnadu, India[1]

Head of the Department of Computer Science, Shri Sakthikailaash women's college, Salem, Tamilnadu, India[2]

ABSTRACT: Recently there has been a growing interest in designing differentially private data mining algorithms. A variety of algorithms have been proposed for mining frequent itemsets. Frequent itemset mining (FIM) is one of the most fundamental problems in data mining. It has practical importance in a wide range of application areas such as decision support, web usage mining, bioinformatics, etc. In this paper, It explore the possibility of designing a differentially private FIM algorithm which can not only achieve high data utility and a high degree of privacy, but also offer high time efficiency. To this end, a differentially private FIM algorithm based on the FP-growth algorithm, which is referred to as PFP-growth. The PFP-growth consist of a preprocessing phase and a mining phase. In the preprocessing phase, to improve the utility and privacy tradeoff, a novel smart splitting method is proposed to transform the database. For a given database, the preprocessing phase needs to be performed only once. In the mining phase, to offset the information loss caused by transaction splitting,we devise a run-time estimation method to estimate the actual support of itemsets in the original database. In addition, by leveraging the downward closure property, we put forward a dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process. Through formal privacy analysis, PFP-growth algorithm is differentially private. Extensive experiments on real datasets illustrate that in PFP-growth algorithm substantially outperforms the state-of-the-art technique.

KEYWORDS: Private FIM algorithm, Smart splitting, Run time Estimation

I. INTRODUCTION

Frequent itemset mining (FIM) is one of the most fundamental problems in data mining. It has practical importance in a wide range of application areas such as decision support, Web usage mining, bioinformatics, etc. Given a database, where each transaction contains a set of items, FIM tries to find itemsets that occur in transactions more frequently than a given threshold. Despite valuable insights the discovery of frequent itemsets can potentially provide, if the data is sensitive (e.g., web browsing history and medical records), releasing the discovered frequent itemsets might pose considerable threats to individual privacy. Differential privacy has been proposed as a way to address such problem. Unlike the anonymization-based privacy models (e.g., k-anonymity and l-diversity), differential privacy offers strong theoretical guarantees on the privacy of released data without making assumptions about an attacker's background knowledge. In particular, by adding a carefully chosen amount of noise, differential privacy assures that the output of a computation is insensitive to changes in any individual's record, and thus restricting privacy leaks through the results. A variety of algorithms have been proposed for mining frequent itemsets. The Apriori and FP-growth are the two most prominent ones. In particular, Apriori is a breadth first search, candidate set generation-and-test algorithm. It needs l database scans if the maximal length of frequent itemsets is l. In contrast, FP-growth is a depth-first search algorithm, which requires no candidate generation. Compared with Apriori, FP-growth only performs two database scans, which makes FP-growth an order of magnitude faster than Apriori. The appealing features of FP-growth motivate us to design a differentially private FIM algorithm based on the FP-growth algorithm. In this paper, we argue that a practical differentially private FIM algorithm should not only achieve high data utility and a high degree of privacy, but also offer high time efficiency. Although several differentially private FIM algorithms have been proposed, we are not aware of any existing studies that can satisfy all these requirements simultaneously. The resulting

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

demands inevitably bring new challenges. It has been shown that the utility-privacy tradeoff can be improved by limiting the length of transactions. Existing work presents an Apriori-based differentially private FIM algorithm. It enforces the limit by truncating transactions (i.e., if a transaction has more items than the limit, deleting items until its length is under the limit). In particular, in each database scan, to preserve more frequency information, it leverages discovered frequent itemsets to re-truncate transactions. However, FP-growth only performs two database scans. There is no opportunity to re-truncate transactions during the mining process. Thus, the transaction truncating approach proposed is not suitable for FP-growth. In addition, to avoid privacy breach, we add noise to the support of itemsets. Given an i-itemset X (i.e., X contains i items), to satisfy differential privacy, the amount of noise added to the support of i-itemset X depends on the number of support computations of i-itemsets. Unlike Apriori, FPgrowth is a depth-first search algorithm. It is hard to obtain the exact number of support computations of i-itemsets during the mining process. A naive approach for computing the noisy support of i-itemset X is to use the number of all possible i-itemsets. However, it will definitely produce invalid results.

II. RELATED WORK

The Apriori and FP-growth are the two most prominent ones. In particular, Apriori is a breadth first search, candidate set generation-and-test algorithm. It needs *l* database scans if the maximal length of frequent itemsets is *l*. FP-growth is a depth-first search algorithm, which requires no candidate generation. Compared with Apriori, FP-growth only performs two database scans, which makes FP-growth an order of magnitude faster than Apriori. Although several differentially private FIM algorithms have been proposed, we are not aware of any existing studies that can satisfy all these requirements simultaneously. Existing work presents an Apriori-based differentially private FIM algorithm. It enforces the limit by truncating transactions (i.e., if a transaction has more items than the limit, deleting items until its length is under the limit). FP-growth only performs two database scans. There is no opportunity to re-truncate transactions during the mining process. Thus, the transaction truncating approach proposed in is not suitable for FP-growth. Unlike Apriori, FPgrowth is a depth-first search algorithm. It is hard to obtain the exact number of support computations of *i*-itemsets during the mining process. it will definitely produce invalid results.

Disadvantages:

- It is hard to obtain the exact number of support computations of *i*-itemsets during the mining process.

It enforces the limit by truncating transactions (i.e., if a transaction has more items than the limit, deleting items until its length is under the limit).

III. PROPOSED SYSTEM

To address these challenges, we present our private FPgrowth (PFP-growth) algorithm, which consists of a preprocessing phase and a mining phase. In the preprocessing phase, we transform the database to limit the length of transactions. The preprocessing phase is irrelevant to user specified thresholds and needs to be performed only once for a given database. We argue, to enforce such a limit, long transactions should be split rather than truncated. That is, if a transaction has more items than the limit, we divide it into multiple subsets (i.e., sub-transactions) and guarantee each subset is under the limit. In the mining phase, given the transformed database and a user-specified threshold, we privately discover frequent itemsets. During the mining process, we dynamically estimate the number of support computations, so that we can gradually reduce the amount of noise required by differential privacy. In the mining phase, to offset the information loss caused by transaction splitting, we devise a run-time estimation method to estimate the actual support of itemsets in the original database. Runtime estimation method to quantify the information loss caused by transaction splitting Dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process. We explore the possibility of designing a differentially private FIM algorithm which can not only achieve high data utility and a high degree of privacy, but also offer high time efficiency.

Advantages:

- PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

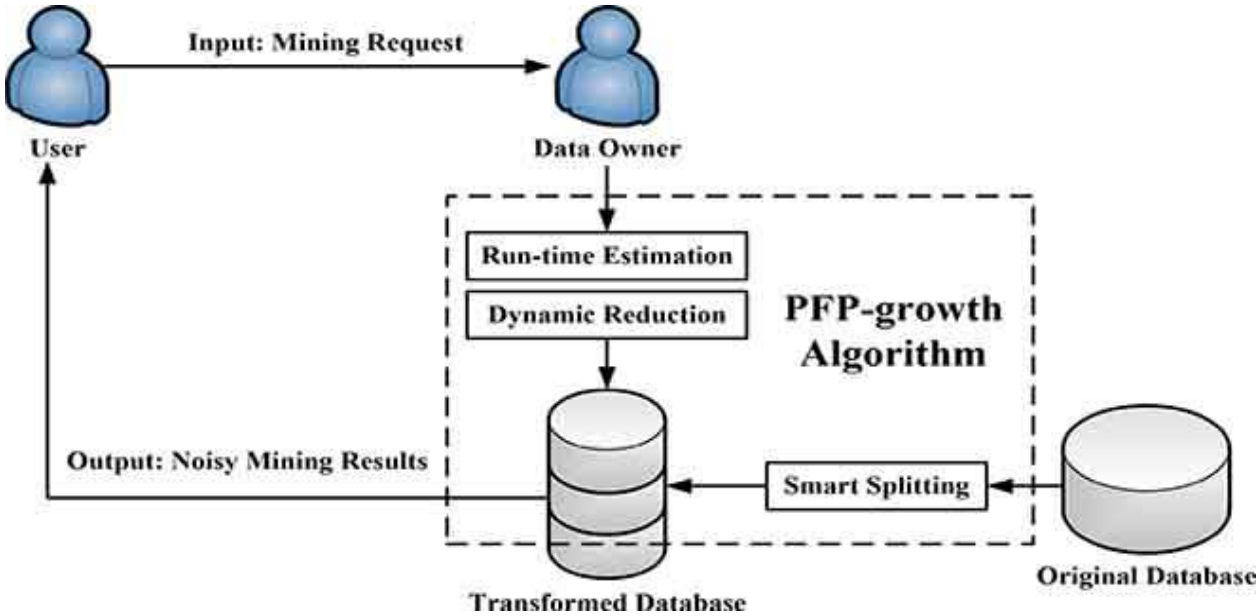
(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

- The preprocessing phase does not consume too much time.

The performance is significantly improved by adopting our transaction Splitting techniques.

IV. SYSTEM ARCHITECTURE



Through formal privacy analysis, we show that our PFPgrowth algorithm is ϵ-differentially private. Extensive experimental results on real datasets show that our algorithm outperforms existing differentially private FIM algorithms. Moreover, to demonstrate the generality of our transaction splitting techniques and further enrich the application spectrum, we apply our transaction splitting techniques, including the smart splitting and run-time estimation methods, to Apriori by modifying the algorithm. Preliminary experimental results show that the performance of the Apriori-based algorithm is significantly improved by adopting our transaction splitting techniques. To summarize, our key contributions are:1). We revisit the tradeoff between utility and privacy in designing a differentially private FIM algorithm. We demonstrate that the tradeoff can be improved by our novel transaction splitting techniques. Such techniques are not only suitable for FP-growth, but also can be utilized to design other differentially private FIM algorithms. 2). We develop a time-efficient differentially private FIM algorithm based on the FP-growth algorithm, which is referred to as PFP-growth. In particular, by leveraging the downward closure property, a dynamic reduction method is proposed to dynamically reduce the amount of noise added to guarantee privacy during the mining process. 3). Through formal privacy analysis, we show that our PFP-growth algorithm is ϵ-differentially private. Extensive experiments on real datasets illustrate our algorithm substantially outperforms the state-of-the-art techniques.

- **PFP-GROWTH ALGORITHM**

The PFP-growth algorithm consists of two phases.

1. In particular, in **the preprocessing phase**, we extract some statistical information from the original database and leverage the smart splitting method to transform the database.
2. Notice that, for a given database, the preprocessing phase is performed only once.

In the **mining phase**, for a given threshold, we privately find frequent itemsets.
The run-time estimation and dynamic reduction methods are used in this phase to improve the quality of the results.
Besides, we divide the total privacy budget ϵ into five portions:
$\epsilon 1$ is used to compute the maximal length constraint,

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

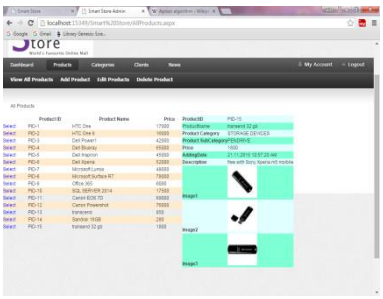
$\epsilon 2$ is used to estimate the maximal length of frequent itemsets,
$\epsilon 3$ is used to reveal the correlation of items within transactions,
$\epsilon 4$ is used to compute μ-vectors of itemsets, and
$\epsilon 5$ is used for the support computations.

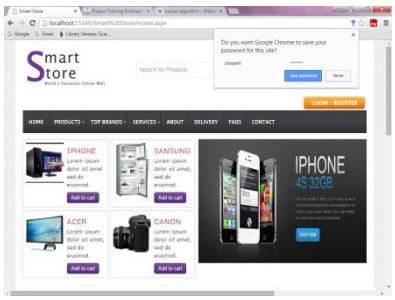
PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy.

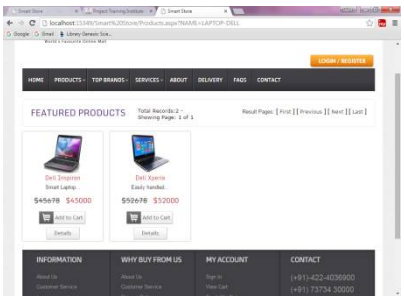
V. SIMULATION AND RESULT

WORK TO BE DONE IN PHASE II:

- The run-time estimation and dynamic reduction methods are used in this phase to improve the quality of the results.
- Runtime estimation method to quantify the information loss caused by transaction splitting.
- Dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process.







VI. CONCLUSION AND FUTUREWORK

We investigate the problem of designing a differentially private FIM algorithm. We propose our private FP-growth (PFP-growth) algorithm, which consists of a preprocessing phase and a mining phase. In the preprocessing phase, to better improve the utility-privacy tradeoff, we devise a smart splitting method to transform the database. In the mining phase, a run-time estimation method is proposed to offset the information loss incurred by transaction splitting. Moreover, by leveraging the downward closure property, we put forward a dynamic reduction method to dynamically reduce the amount of noise added to guarantee privacy during the mining process. Formal privacy analysis and the results of extensive experiments on real datasets show that our PFP-growth algorithm is time-efficient and can achieve both good utility and good privacy. Further reduce the sensitivity while avoiding too much overhead is an interesting direction for future work.

REFERENCES

[1]. Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Yang, "Differentially Private Frequent Itemset Mining via Transaction Splitting"in Knowledge and Data Engineering 2015.
[2].L. Bonomi and L. Xiong, "A two-phase algorithm for mining sequential patterns with differential privacy," in CIKM, 2013.
[3]. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in KDD, 2002.
[4].W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "Security in outsourcing of association rule mining," in VLDB,2007.
[5]. W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "An audit environment for outsourcing of frequent itemset mining," in VLDB, 2009.
[6]. R. Chen, B. C. M. Fung, and B. C. Desai, "Differentially private transit data publication: A case study on the montreal transportation system," in KDD, 2012.
[7] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in CCS, 2012.
[8] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," SIAM Journal on Computing, 2012.
[9] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," SIGKDD Explorations, 2004.
[10] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," J. Statistical Mechanics:

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

Theory and Experiment, 2008.

BIOGRAPHY

A.Monisha is a M.Phil student in department of computer science, Shri sakthikailaash women's college, Salem, India. She received her M.Sc computer science degree in from bharathidhasan university, Trichy, India. She research interests are datamining and frequent itemset mining.

B.S.Sangeetha M.C.A..M.Phil..B.Ed. is a Head of the Department of Computer Science in Shri Sakthikailaash Women's college, Salem.She research interest are data mining, frequent itemset mining and networking.